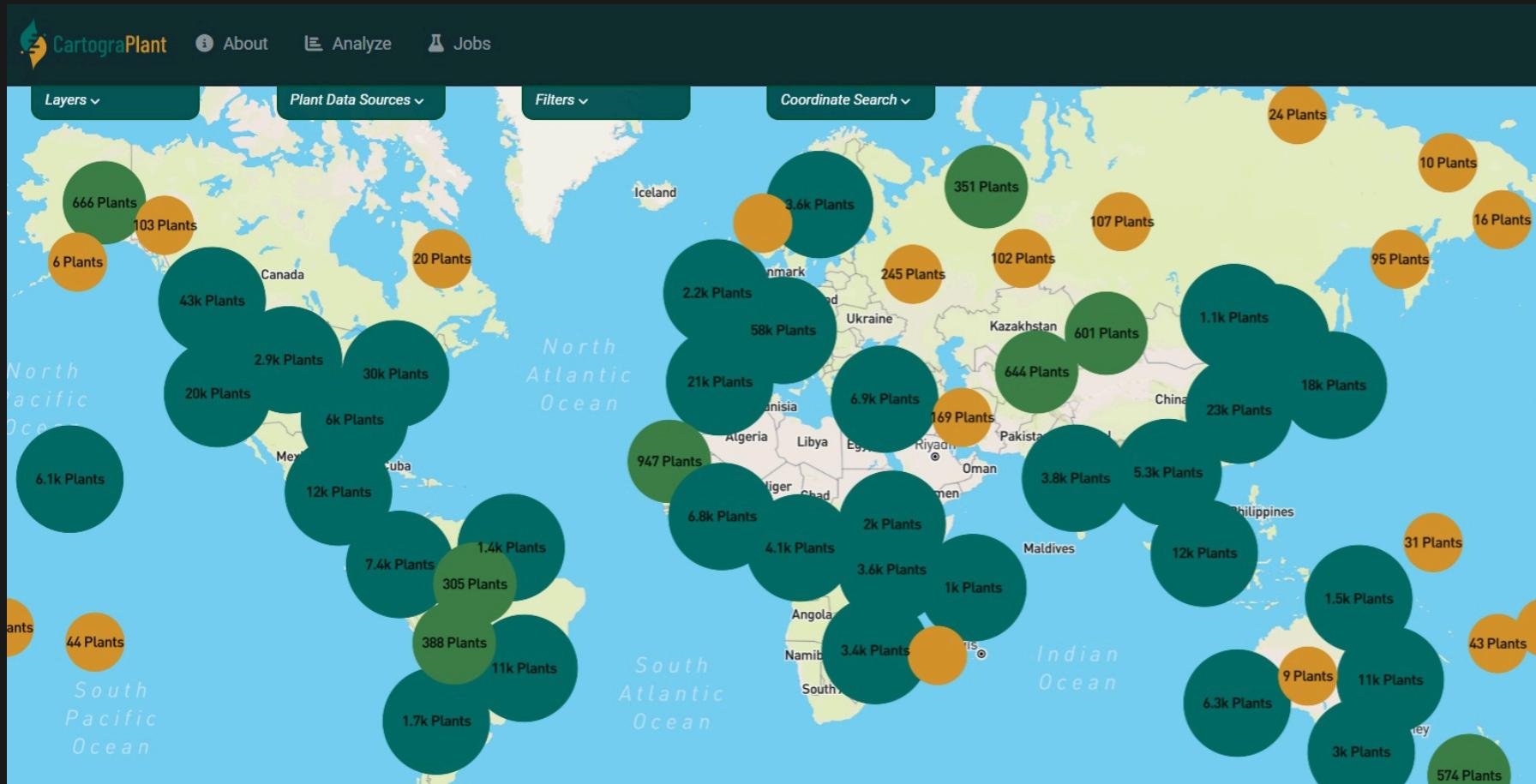
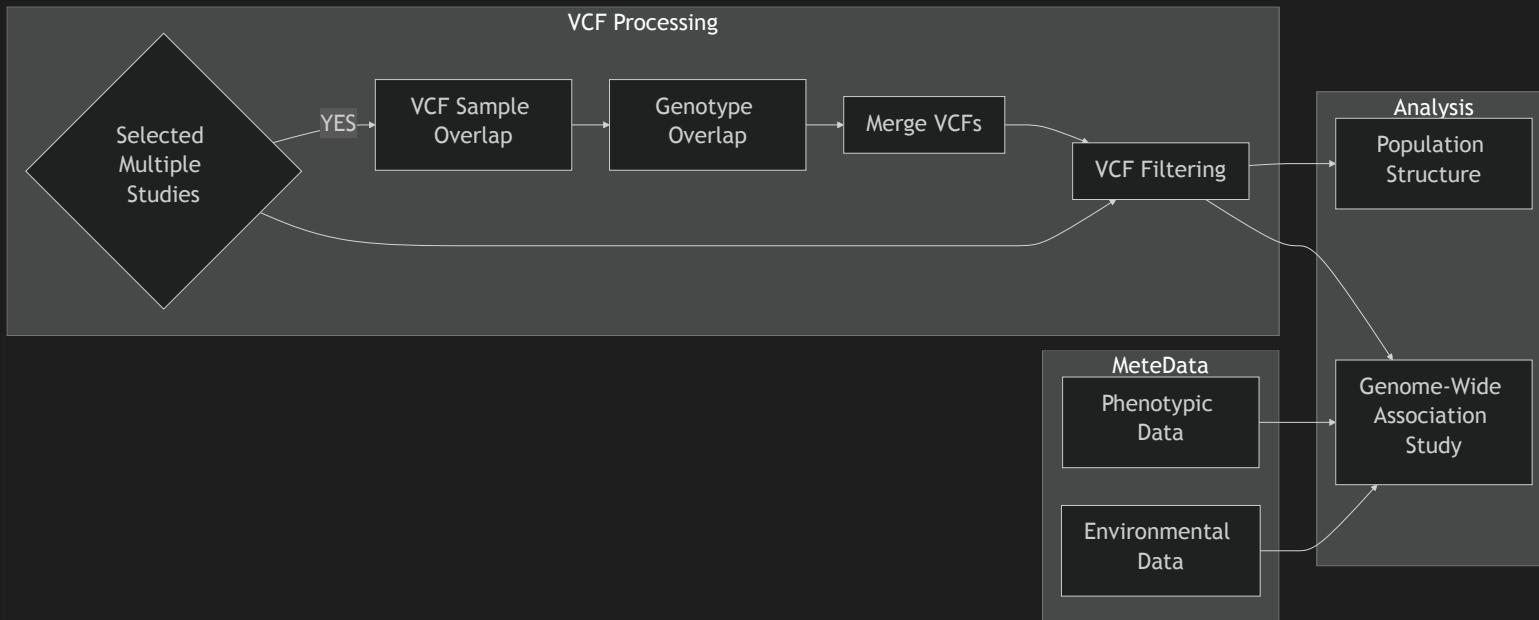


| Analysis Panel | Landscape Genomics | Demo |



ANALYSIS PANEL OVERVIEW



NEXTFLOW

Workflow management software



Reproducible



Concurrency



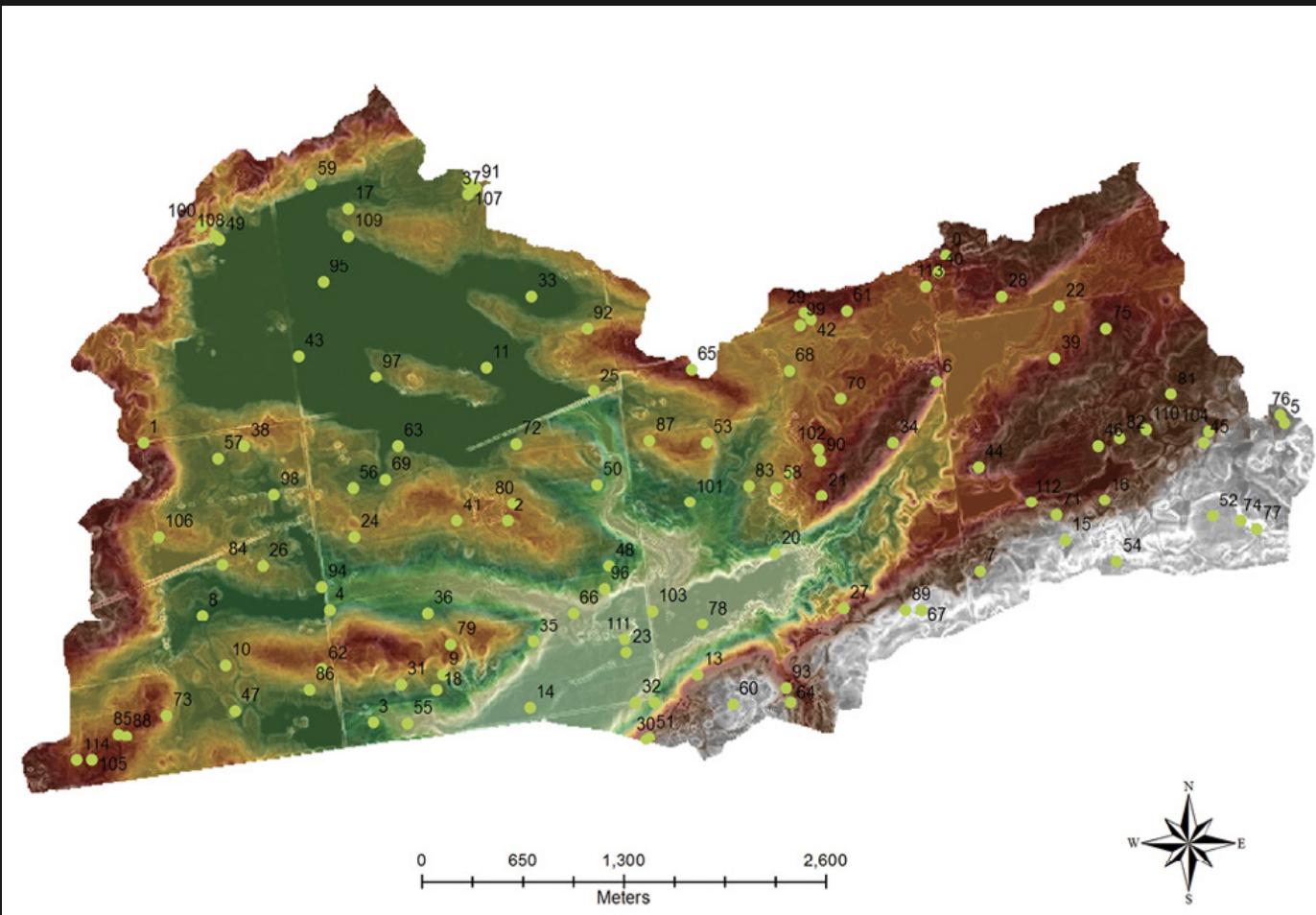
Portable



Community

LANDSCAPE GENOMICS

Aims to understand how environmental heterogeneity influences the distribution of genetic diversity.



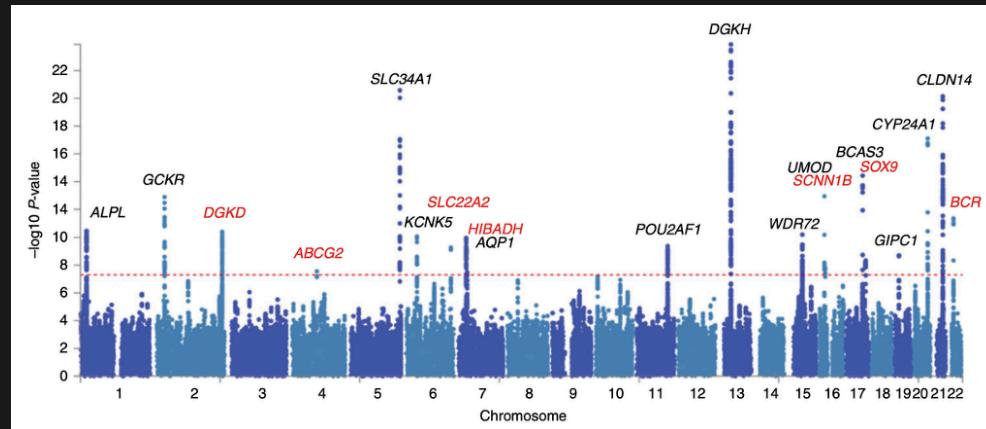
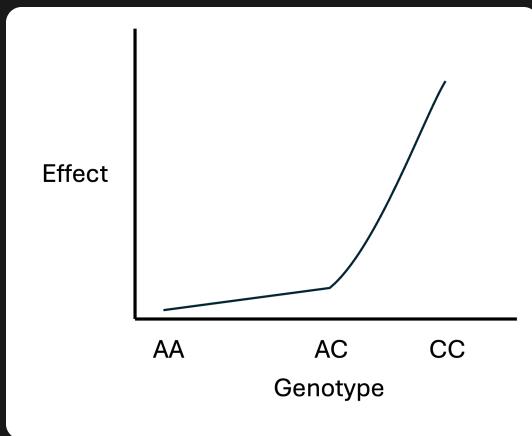
POPULATION STRUCTURE

Patterns of genetic variation within and between individuals

- Isolation by distance
- Genetic drift
- Selection

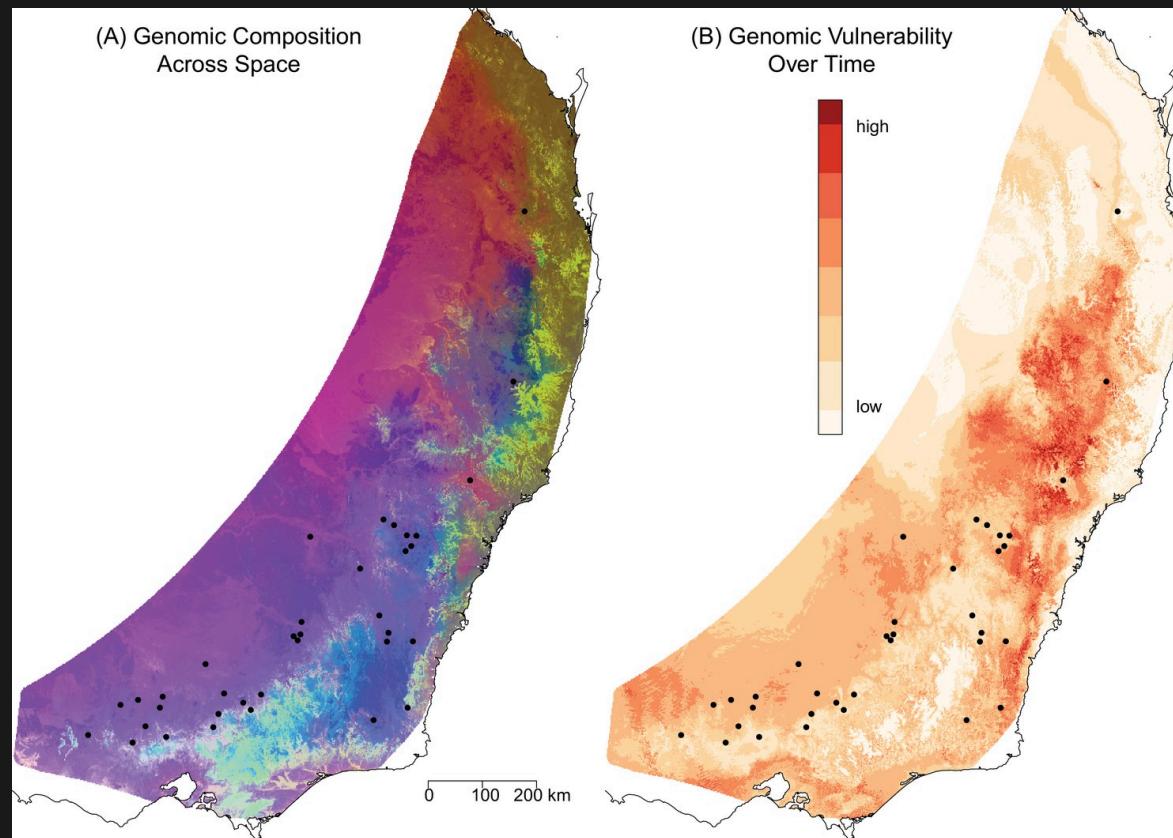
GENOME-WIDE ASSOCIATION STUDY (GWAS)

Examines the relationship between different genotypes and specific traits or environments.



GENETIC OFFSET

Difference between the current genetic makeup to the genetic makeup predicted to be optimal for future environmental conditions



TPPS SUBMISSION ENABLES SEAMLESS INTEGRATION OF STUDIES

Allowing for analysis of large datasets in the analysis panel

Mega:
Integration of
raw data

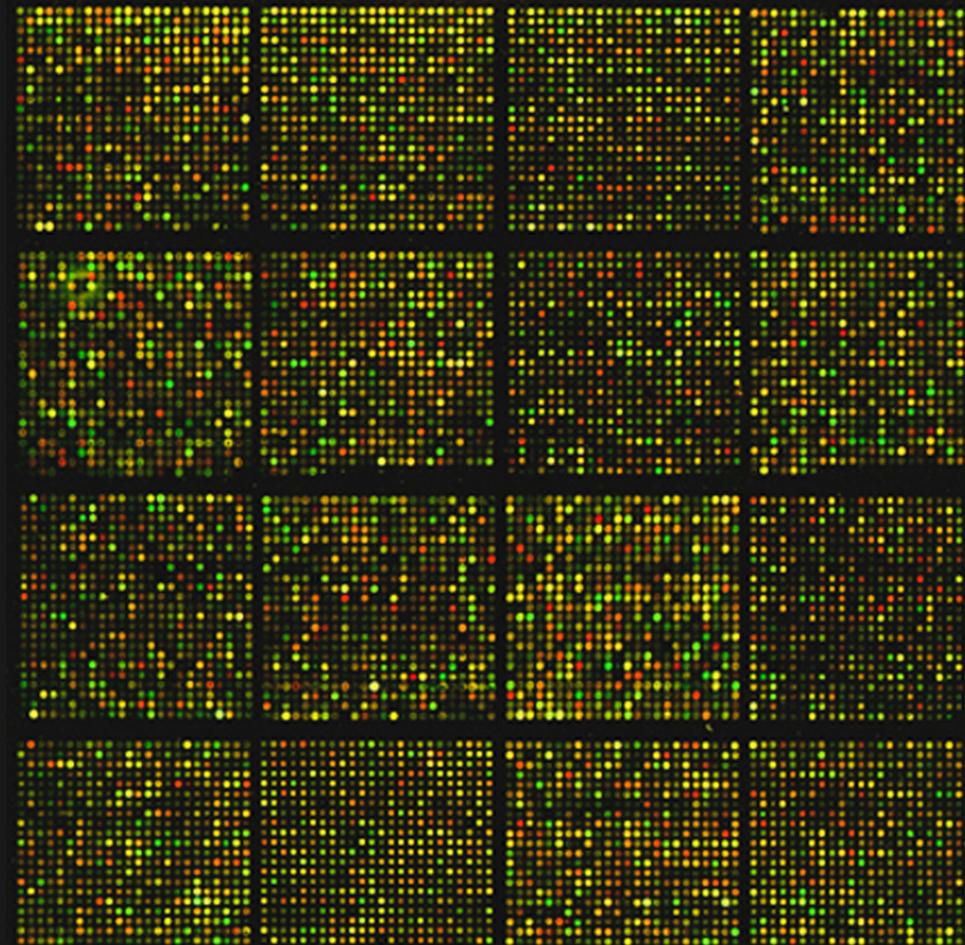
Meta:
Integration of
summary
results

DEMO WITH *POPULUS TRICHOCARPA*

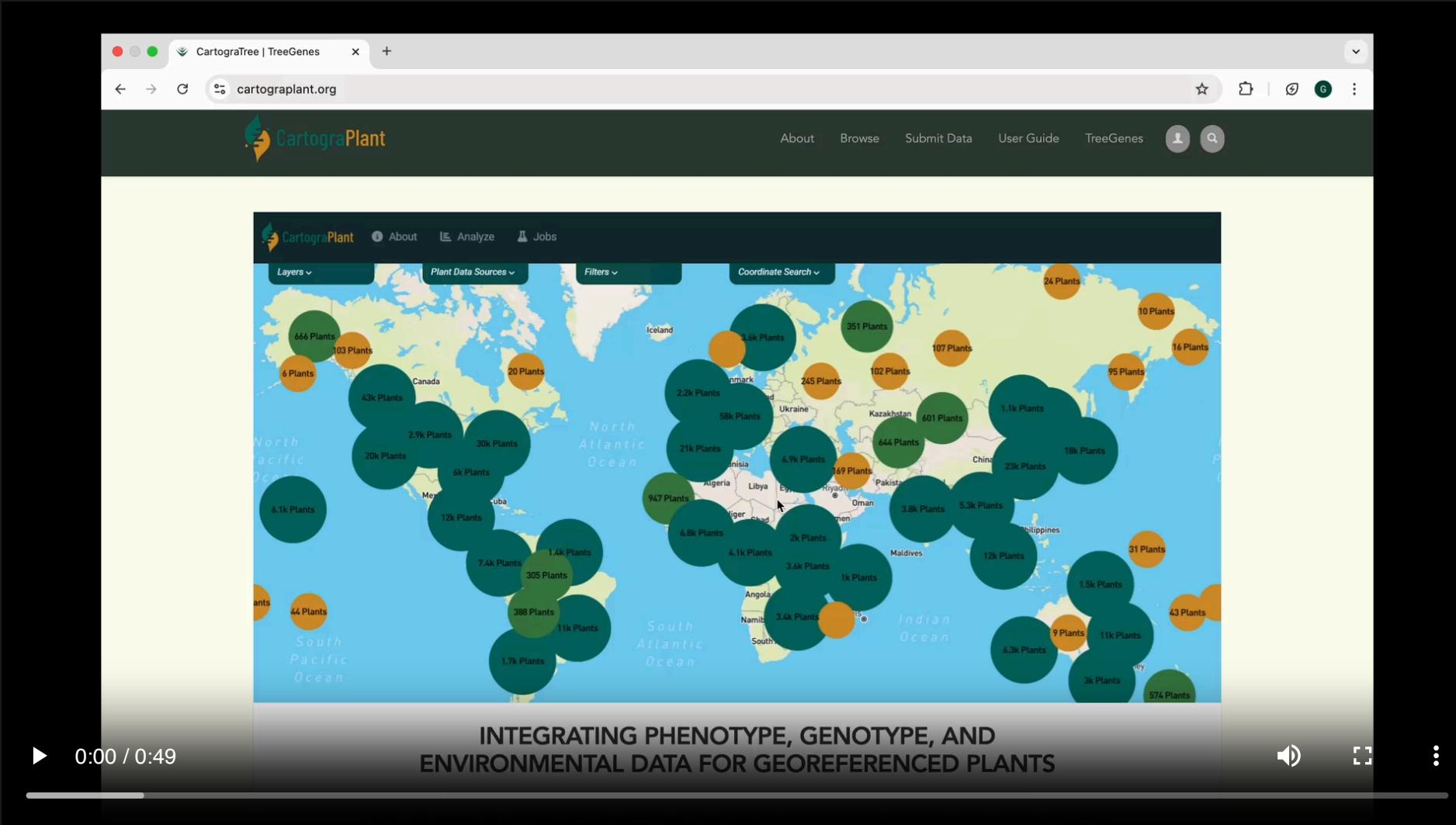
| Study | Samples | Genotypes | Phenotypes |
|-------------------------|---------|-----------|------------|
| Geraldes et al. 2008 | 55 | 32,000 | 0 |
| Mckown et al. 2013 | 555 | 0 | 400 |
| Mckown et al. 2014 | 555 | 28,000 | 0 |

GENOTYPE ASSAY

A cost effective way to obtain tens of thousands of mutations



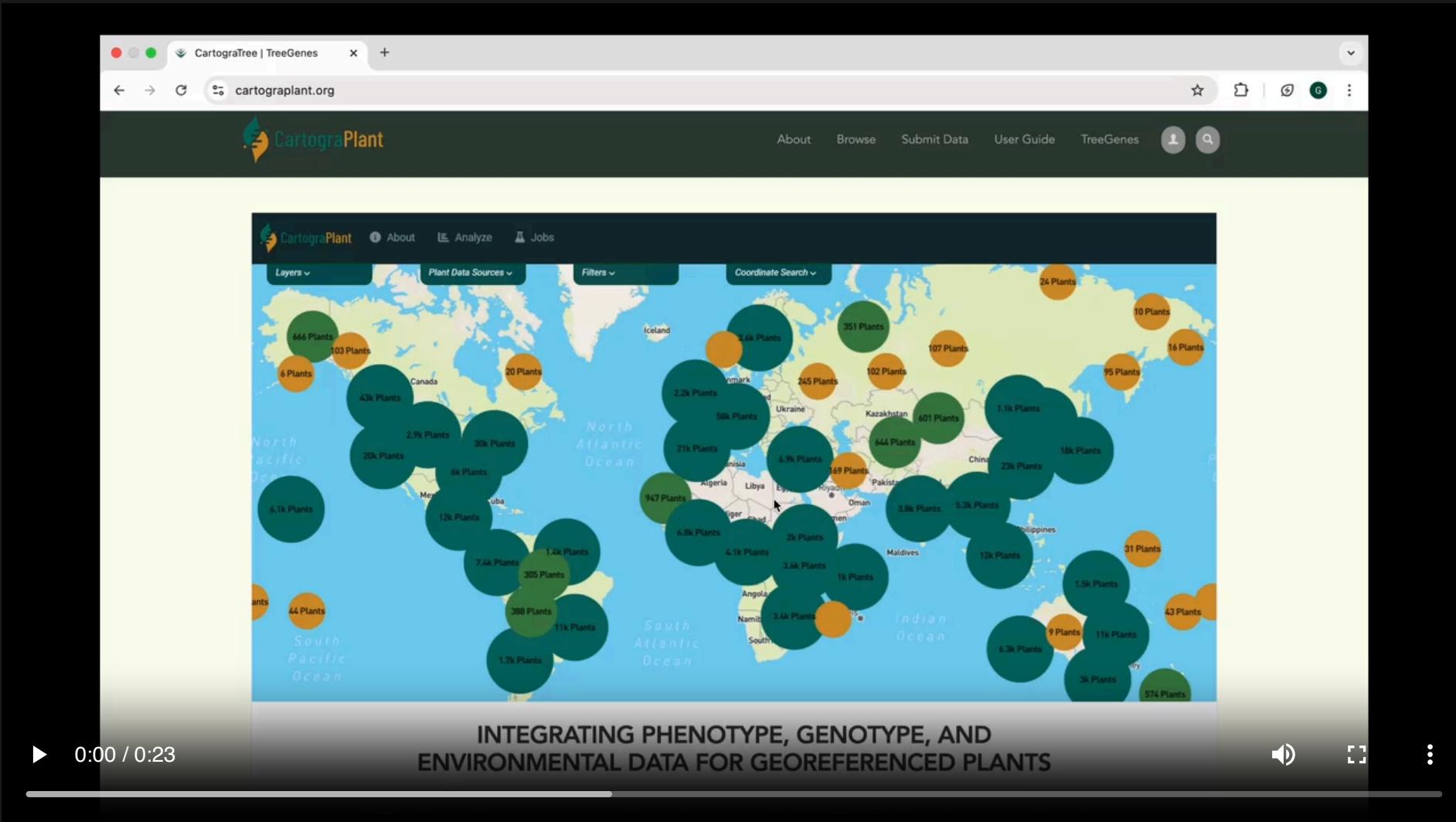
FILTER AND SELECT STUDIES IN CARTOGRAPLANT



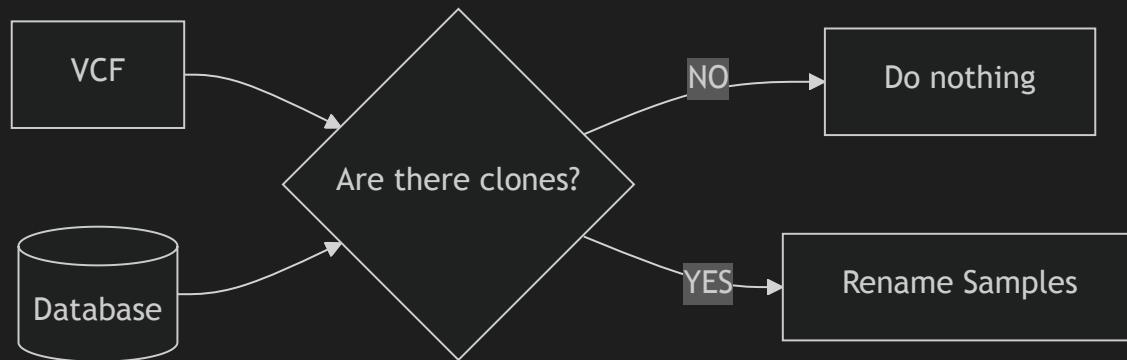
PANEL 1: WORKSPACE

Responsible for file management:

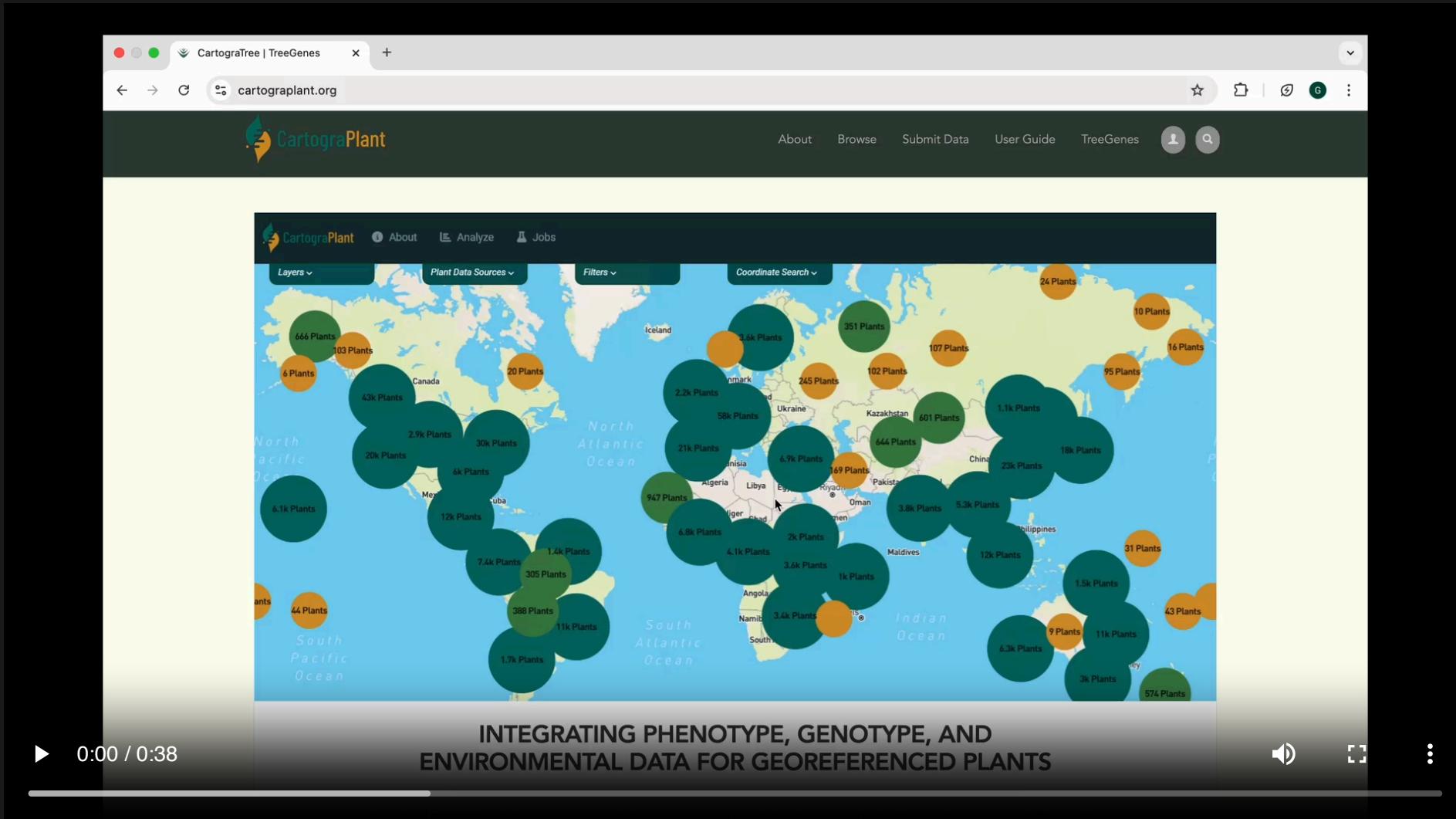
- Create new workspace directory
- Upload files from local machine
- Stores output from workflows
- Delete Files



PANEL 2: STUDY-CONTEXT



| Study 1 | Study 2 | Study 3 | Standard Name |
|-----------|----------|----------|---------------|
| GLCA-26-1 | GLCA26-1 | GLCA26-1 | sample05 |
| HFCA-20-2 | Null | Null | sample06 |



PANEL 3: FILTER BY TRAITS

Select phenotypes of interest for upload into workspace

- Determine total counts
- Redundancy analysis based on $x > 1$ phenotypes

CartograTree | TreeGenes

treegenesdb.org/cartogratree#

CartograPlant

Analysis ID: 3541 | Studies: 3 | Plants: 963 | Species: 1 | Phenotypes: 2 | Genotypes: 28083 | Environmental layers: 2

Manage analysis

Study context

Filter By Traits

Study markers overlap

Filtering & Imputation

Population Structure

Environmental Data

Run Analysis

Summary and Confirm

3 studies based on analysis study context choices

TGDR1892 | TGDR1904 | TGDR2269

| | | | |
|---|------------------|--|-------------|
| <input type="checkbox"/> bole density | 519 phenotypes | | No overlaps |
| <input type="checkbox"/> bole mass | 519 phenotypes | | No overlaps |
| <input type="checkbox"/> bud break | 916 phenotypes | | No overlaps |
| <input checked="" type="checkbox"/> bud set | 1,370 phenotypes | | No overlaps |
| <input type="checkbox"/> canopy duration | 910 phenotypes | | No overlaps |
| <input checked="" type="checkbox"/> change in plant height | 1,368 phenotypes | | No overlaps |
| <input type="checkbox"/> change in plant volume | 916 phenotypes | | No overlaps |
| <input type="checkbox"/> day 100% of leaves are yellow | 455 phenotypes | | No overlaps |
| <input type="checkbox"/> day 25% of leaves are yellow | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> day 50% of leaves are yellow | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> day 75% of leaves are yellow | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> elevation | 443 phenotypes | | No overlaps |
| <input type="checkbox"/> frost free days | 443 phenotypes | | No overlaps |
| <input type="checkbox"/> leaf chlorophyll content, spring | 203 phenotypes | | No overlaps |
| <input type="checkbox"/> leaf chlorophyll content, summer | 777 phenotypes | | No overlaps |
| <input type="checkbox"/> leaf flush | 1,571 phenotypes | | No overlaps |
| <input type="checkbox"/> leaf lifespan | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> leaf shape | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> log(plant height growth) | 457 phenotypes | | No overlaps |
| <input type="checkbox"/> log(plant volume growth) | 457 phenotypes | | No overlaps |
| <input type="checkbox"/> number of branches | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> plant height | 1,826 phenotypes | | No overlaps |
| <input type="checkbox"/> plant height growth cessation date | 458 phenotypes | | No overlaps |
| <input type="checkbox"/> whole plant mass | 519 phenotypes | | No overlaps |

PCA (2 phenotypes)

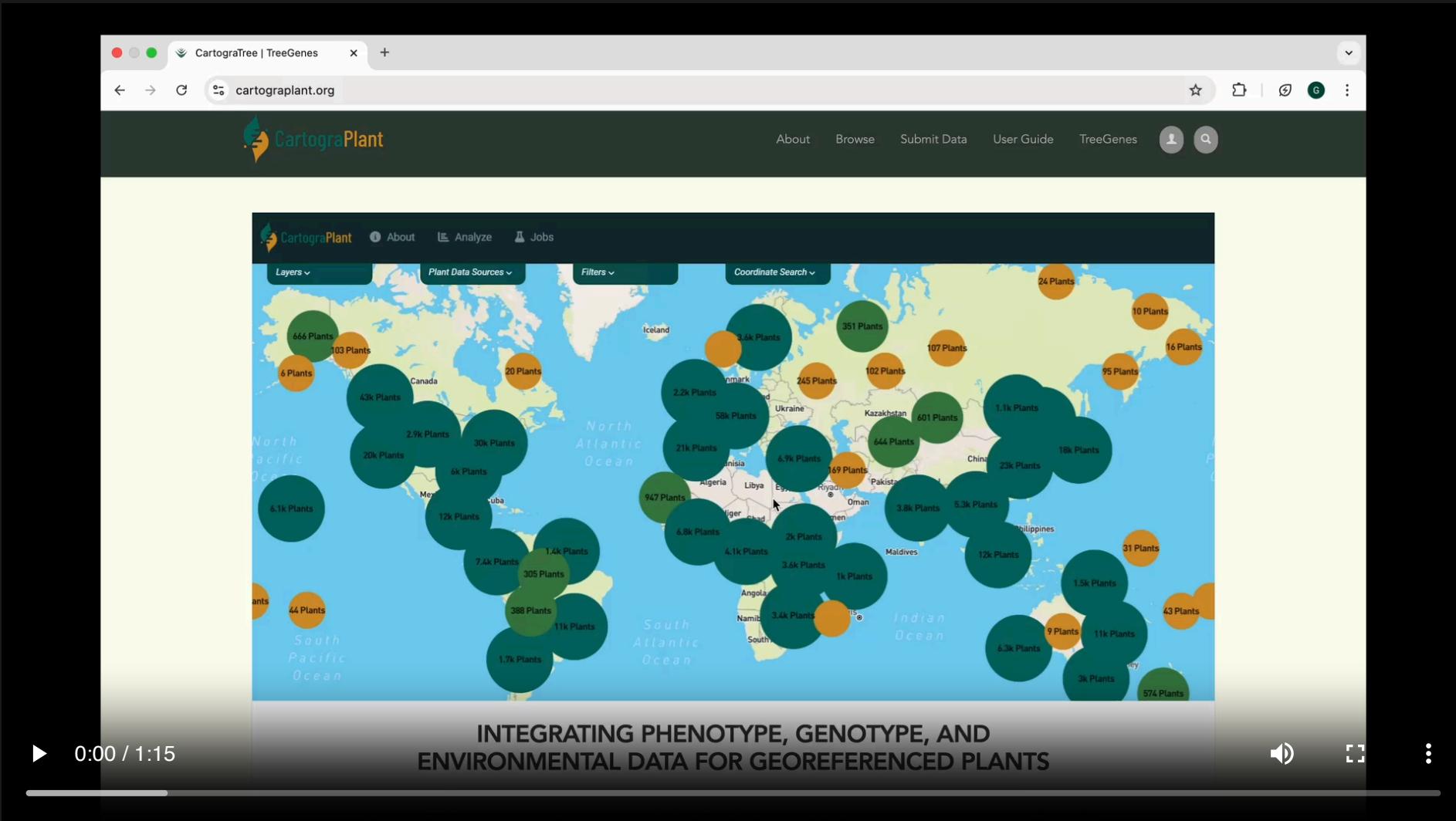
PHENOTYPE FILE OUTPUT FORMAT

| phenotype_id | phenotype_name | plant_accession | study_accession | year | value |
|--------------|------------------------|-------------------|-----------------|------|-------|
| 8880144 | change in plant height | TGDR1904-ALAA20-1 | TGDR1904 | 2009 | 194.7 |

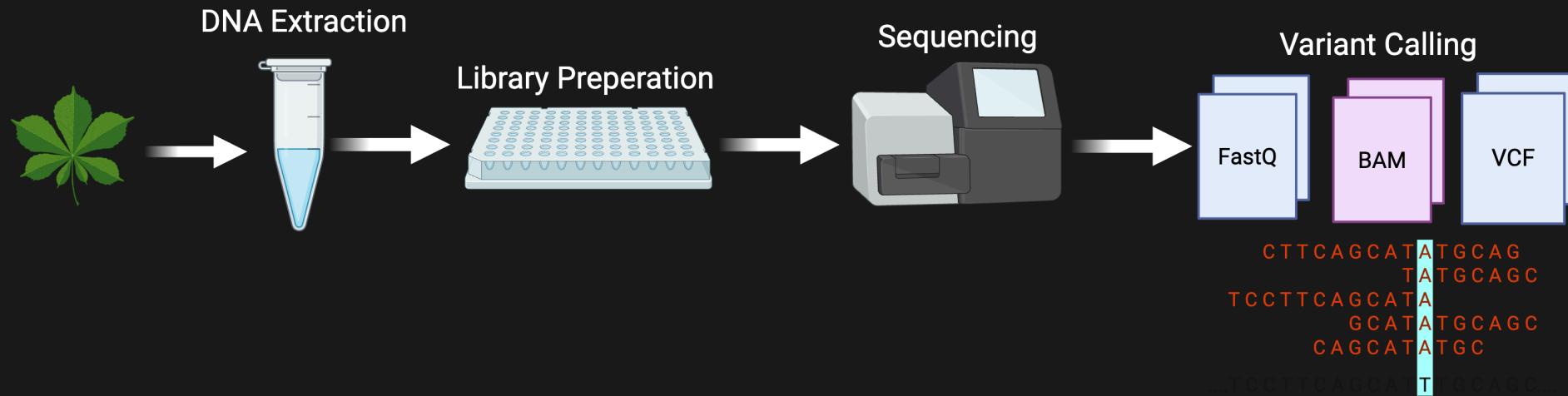
PANEL 4: STUDY MARKERS OVERLAP

1. Marker Overlap
2. Genotype concordance
3. Merge VCF's

GENOTYPE QUALITY CONTROL



PANEL 5: VARIANT FILTERING



CartograTree | TreeGenes

treegenesdb.org/cartogratree#

Dashboard Content Structure Tripal Appearance People Modules Mainlab CartograTree Admin Configuration TG Gus Reports Help Hello gabriel.barrett@uconn.edu Log out

Analysis ID: 3536 Studies: 3 Plants: 963 Species: 1 Phenotypes: 0 Genotypes: 28083 Environmental layers: 0

Manage analysis Study context Filter By Traits Study markers overlap Filtering & Imputation Population Structure Environmental Data Run Analysis Summary and Confirm

Variant filtering - TGDR1892

Filter and Imputation

💡 Filtering variables retrieved, rules can now be created.

Statistics variables

OriginalAlleles OriginalContig OriginalStart ReverseComplementedAlleles
 SwappedAlleles GT

Filter and imputation rules

AND OR ADD RULE ADD GROUP

INFO ✓ equal greater less A list of the original alleles (includi... DELETE

value

Variant filtering - TGDR2269

Filter and Imputation

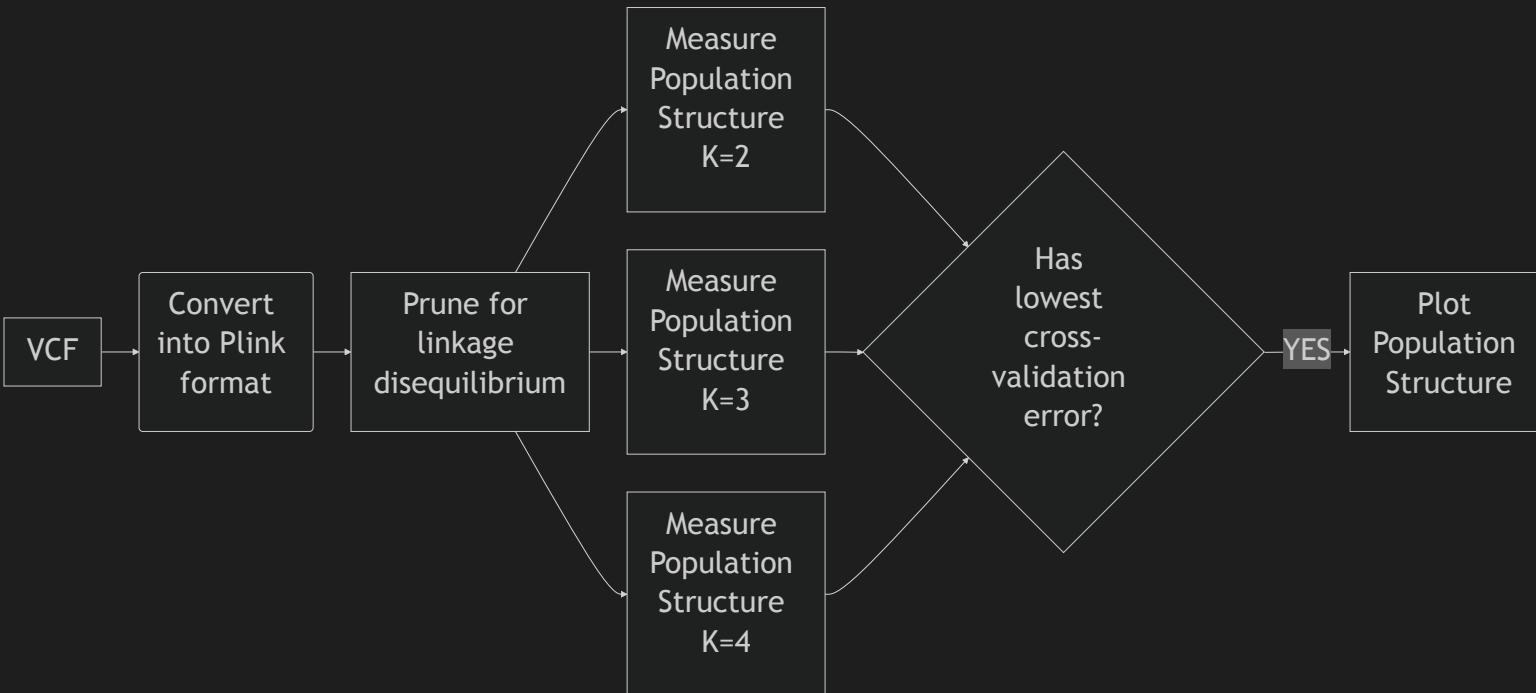
💡 Filtering variables retrieved, rules can now be created.

Statistics variables

OriginalAlleles OriginalContig OriginalStart ReverseComplementedAlleles
 SwappedAlleles GT

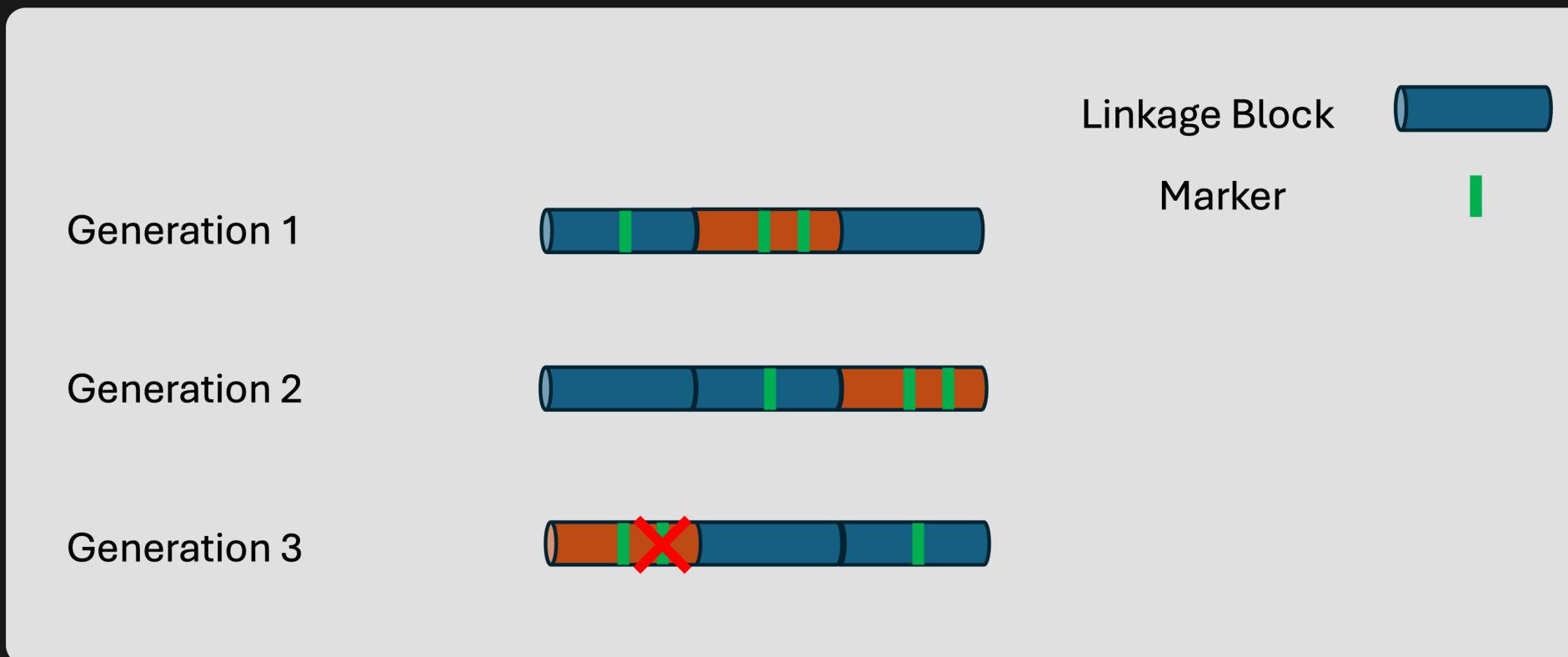
Filter and imputation rules

PANEL 6: POPULATION STRUCTURE



LINKAGE DISEQUILIBRIUM

Non-random associate of alleles at different loci on a chromosome



CartograTree | TreeGenes

treegenesdb.org/cartogratree#

Dashboard Content Structure Tripal Appearance People Modules Mainlab CartograTree Admin Configuration TG Gus Reports Help Hello gabriel.barrett@uconn.edu Log out

Analysis ID 3526 Studies 3 Plants 963 Species 2 Phenotypes 0 Genotypes 28083 Environmental layers 2

Manage analysis Study context Study markers overlap Filtering & Imputation Population Structure Environmental Data Run Analysis Summary and Confirm

Nextflow Population Structure Workflow

Path To Vcf File
Geraldes_2013.v2_TGDR1892_Potr.2_0-Potr.4_1.vcf.vcf.gz

Select appropriate variant call file

Model Complexity
3,4,5

Window Size To Calculate Linkage Disequilibrium
150

Number Of Variants To Consider In A Window
10

The Correlation Between A Pair Of Loci
0.2

How To Handle Partially Missing Genotypes When Converting File
Treat half-calls as haploid/homozygous

How To Handle Converting VCF ID's To Plink IID And FIID
0

GENERATE POPULATION STRUCTURE

NEXT CLOSE

K=4

Ancestry

Location

https://treegenesdb.org/cartogratree#analysis-popstruct-section

PANEL 7: CHOOSE ENVIRONMENTS

multicollinearity: describes a high correlation of two or more independent variables

CartograTree | TreeGenes x

treegenesdb.org/cartogratree#

Dashboard Content Structure Tripal Appearance People Modules Mainlab CartograTree Admin Configuration TG Gus Reports Help Hello gabriel.barrett@uconn.edu Log out

CartograPlant

Analysis ID: 3541 Studies: 3 Plants: 963 Species: 1 Phenotypes: 0 Genotypes: 0 Environmental layers: 1

Manage analysis Study context Filter By Traits Study markers overlap Filtering & Imputation Population Structure **Environmental Data** Run Analysis Summary and Confirm

Choose environmental layers

CATEGORY US

GROUP Biotic Damage (North America)

GROUP Forest Fragmentation (North America, ESRI)

GROUP Climatic variables (World, ClimateWNA)

SEARCH frost 1 matches found.

SELECT ALL

CLASSIFICATION PERIOD

PROPERTY Frost-free period CWNA Variable FFP

LAYER Mean coldest monthly temperature CWNA-BC - Celsius

LAYER Hargreaves reference evaporation CWNA

GROUP NDVI (USGS)

GROUP Climatic variables (World, AWI/MARUM)

GROUP Drainage (US, USGS)

CATEGORY WORLD

GROUP Climatic variables (World, WorldClim v.2)

By adjusting your selected environmental layers, the PCA scatterplot will (re)generate after you click Gather and upload to workspace button.

PRECACHE VALUES

GATHER AND UPLOAD TO WORKSPACE

NEXT

CLOSE

MINN
APOLIS
IOWA
Cheyenne
Omaha

PANEL 8: RUN ANALYSIS

Perform G x P and G x E associations

LINEAR MODEL (LM) IN GEMMA

CartograTree | TreeGenes

treegenesdb.org/cartogratree#

Manage analysis

Study context

Filter By Traits

Study markers overlap

Filtering & Imputation

Population Structure

Environmental Data

Run Analysis

Summary and Confirm

Nextflow - GWAS Workflow

Path To Vcf File

McKown_2014_GWAS.v3_TGDR2269_Potr.3_0-Potr.4_1.vcf.vcf.gz

Path To Metadata File

1750603468-AN3541_PHENOVER_change in plant height_all-overlaps.csv

Program Type To Run

gemma

Models available

Choose Sub Workflow linear model

Model options

LINEAR MODEL

Step 1 - GEMMA_LM

Type Of Frequentists Test To Perform.

Wald test

likelihood ratio test

score test

all

RUN STEP

GWAS Step Completed

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

0 10 20

SPURIOUS CORRELATION

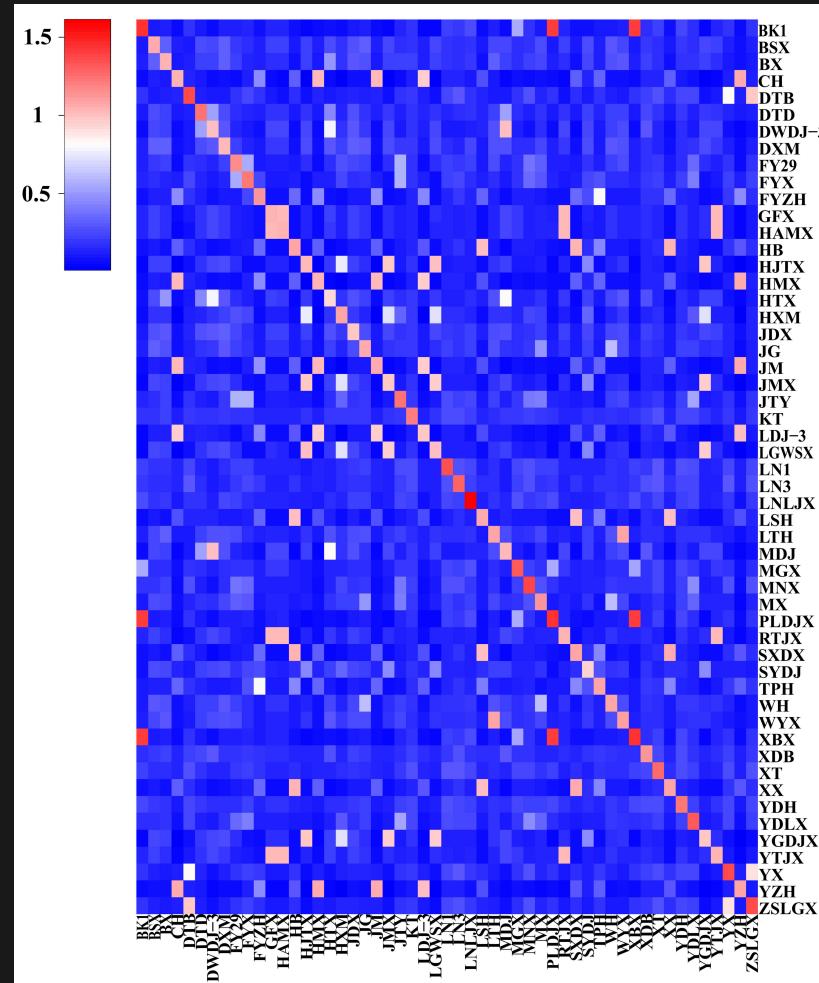
LINEAR MIXED MODEL (LMM) IN GEMMA

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

- \mathbf{y} : vector of phenotypes
- \mathbf{X} : matrix of fixed effects (e.g., SNPs, covariates)
- $\boldsymbol{\beta}$: vector of fixed-effect coefficients
- \mathbf{Z} : incidence matrix for random effects
- $\mathbf{u} \sim \mathcal{N}(0, \mathbf{K}\sigma_g^2)$: random genetic effects with kinship matrix \mathbf{K}
- $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}\sigma_e^2)$: residual errors

KINSHIP MATRIX

Measure the degree of relatedness between samples



LINEAR MIXED MODEL (LMM) IN GEMMA

The screenshot shows a web-based interface for running a GWAS workflow using Nextflow. The main panel is titled "Nextflow - GWAS Workflow". On the left, there's a sidebar with various analysis options like "Manage analysis", "Study context", "Filter By Traits", etc., and a prominent green "Run Analysis" button.

The main configuration area includes:

- Path To Vcf File:** A dropdown menu set to "combined.vcf.gz".
- Path To Metadata File:** A dropdown menu set to "1750543981-AN3537_ENVDATA_Frost-free period CWNA (FFP).csv".
- Program Type To Run:** A dropdown menu set to "gemma".
- Models available:** A dropdown menu set to "linear mixed model".
- Model options:**
 - Step 1 - GEMMA_RELATEDNESS:** A dropdown menu set to "centered matrix".
 - RUN STEP** (A yellow button).
 - Step 2 - GEMMA_LMM:** A dropdown menu set to "Geraldes_2013.v2_TGDR1892_Potr.2_0-Potr.4_1.vcf.vcf_recode.cXX.txt".
 - Type Of Frequentists Test To Perform:** A dropdown menu set to "Wald test".
 - Minimal Value For Lambda:** An input field containing ".00005".
 - Maximum Value For Lambda:** An input field containing "50000".
 - Number Of Regions Used To Evaluate Lambda:** An input field containing "10".
 - RUN STEP** (A yellow button).

At the bottom, a status message reads "Performing GWAS processing...".

LINEAR MIXED MODEL (LMM) IN LEA

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{V}^\top + \boldsymbol{\epsilon}$$

- \mathbf{Y} : matrix of genotypes (individuals \times loci)
- \mathbf{X} : matrix of environmental variables (individuals \times predictors)
- $\boldsymbol{\beta}$: matrix of effect sizes (predictors \times loci)
- \mathbf{U} : matrix of latent factors (individuals \times K)
- \mathbf{V} : matrix of factor loadings (loci \times K)
- $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$: residual noise

LINEAR MIXED MODEL (LMM) IN LEA

The screenshot shows the CartograPlant GWAS Workflow interface on a web browser. The top navigation bar includes links for Dashboard, Content, Structure, Tripal, Appearance, People, Modules, Mainlab, CartograTree Admin, Configuration, TG Gus, Reports, and Help. A user is logged in as gabriel.barrett@uconn.edu.

The main header displays analysis statistics: Analysis ID 3545, Studies 3, Plants 963, Species 1, Phenotypes 0, Genotypes 0, and Environmental layers 1.

The left sidebar lists various analysis options: Manage analysis, Study context, Filter By Traits, Study markers overlap, Filtering & Imputation, Population Structure, Environmental Data, Run Analysis (highlighted in green), and Summary and Confirm.

The central workspace is titled "Nextflow - GWAS Workflow". It contains the following fields:

- Path To Vcf File: Geraldes_2013.v2_TGDR1892_Potr.2_0-Potr.4_1.vcf.vcf.gz
- Path To Metadata File: 1750543981-AN3537_ENVDATA_Frost-free period CWNA (FFP).csv
- Program Type To Run: lea
- Models available: Choose Sub Workflow linear mixed model
- Model options:
 - Step 1 - LEA_LMM
 - Number Of Latent Factors: 4
 - Use The First Principal Component For Association: True
- Buttons: RUN STEP, NEXT, and CLOSE.

A status message at the bottom indicates "Performing GWAS processing...".

Acknowledgements

Biocuration team

Isabella Harding
Trang Nguyen
Phoebe Zhou

Development team

Emily Grau
Gabe Barrett
Risharde Ramanth
Vlad Savitsky

Dr Jill Wegrzyn
Dr Margaret Staton
Dr Stephen Flickin
Dr Nic Herndon



TreeGenes

A Forest Tree Genome Database
treegenesdb.org



CartograPlant

cartoplast.org

UCONN

UNIVERSITY OF CONNECTICUT



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



ECU



PLANT COMPUTATIONAL GENOMICS



treesnap.org



nextflow

The Nature Conservancy The logo for The Nature Conservancy, featuring a green leaf-like shape inside a circle.

